European Research Studies,
Volume XI, Special Issue (3-4) 2007

Methodological Aspects in Using Pearson Coefficient in Analyzing Social and Economical Phenomena

By
Daniela-Emanuela Dănăcică[1]
Ana-Gabriela Babucea[2]

Abstract:

The authors illustrate in this paper a series of methodological aspects generated by the use of Pearson correlation coefficient in analyzing social and economical phenomena. Pearson correlation coefficient is largely used in economics and social sciences; however, the diversified nature and subtle nuances of this concept raises significant methodological issues. This article deals with aspects concerning the factors that impact on the size and interpretation of Pearson correlation coefficient, as well as special cases of this coefficient.

Keywords: Peasron correlation, tetrachoric correlation coefficient.

[1] A*ssistant Professor, Faculty of Economics, Constantin Brâncuşi University of Târgu-Jiu*
[2]*Professor, Faculty of Economics, Constantin Brâncuşi University of Târgu-Jiu*

1. Introduction

In the socio-economic field phenomenon variability is not exclusively determined by the action of a singular factor, but in most cases it is the result of the action of a pluralism of factors. This gives us the possibility, that by the means of the known variation of a factor, to determine the level of another variable it is in certain dependence with.

The causality ratios between social and economical phenomena can be quantified, analyzed and interpreted by the means of correlation and regression analysis. Within it there is studied the dependence between a resultative variable (characteristic), usually marked with Y and one or more independent variables (characteristics), marked with X ($X_1, X_2 \ldots\ldots X_n$ - if there are several factorial variables).

In order to counteract the severe limitations of covariance in the survey of the intensity of the relationship between two variables, a new indicator has been defined, Pearson linear correlation coefficient:

$$r = \frac{\sum_i^n (\frac{x_i - \bar{x}}{\sigma_x})(\frac{y_i - \bar{y}}{\sigma_y})}{n} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \quad (1)$$

The following equation is largely used in practice:

$$r_{y/x} = \frac{n\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{\sqrt{[n\sum_i^n x^2{}_i - (\sum_i^n x_i)^2][n\sum_i^n y_i - (\sum_i^n y_i)^2]}} \quad (2)$$

Pearson linear correlation coefficient is a symmetric measurement, the equality $r_{y/x} = r_{y/x.}$ being checked for the same indicator. It is invariant to data conversion, to the origin and units change of data series.[3] We shall distinguish the following cases:

- If r has the sign « + » and $\beta_1 > 0$ the relationship between the two variables is direct;
- If $r = 0$ and $\beta_1 = 0$, there is no linear dependence between variables, the regression model coincides with a parallel straight line with the axis *Ox*;
- If $r < 0$, then $\beta_1 < 0$, therefore the relationship between the two variables is reversed.

The more r gets values close to +1 or –1, the stronger the correlation between the two variables; the more its values are closer to 0, the weaker the correlation intensity. If $r_{y/x} = 0$, the analyzed variables are independent.

The interval $[-1,+1]$ is divided in practice as follows:

- If $0 \le r_{y/x} \le 0.2$ there is no significant relationship between the analyzed variables,
- If $0.2 \le r_{y/x} < 0.5$ then the relationship between variables is weak;
- If $0.5 \le r_{y/x} < 0.75$ we have an average intensity relationship
- If $0.75 \le r_{y/x} < 0.95$ we have a strong relationship
- If $0.95 \le r_{y/x} \le 1$ between the two variables there is a functional type relationship.

If the individual values of the correlative pair variables are shown as a bi-dimensional table, Pearson linear correlation coefficient shall be calculated according to the formula:

$$r_{y/x} = \frac{\sum_{i,j}^{n,m} (x_i - \bar{x})(y_i - \bar{y}) f_{ij}}{n\sigma_x \sigma_y} \tag{3}$$

Pearson linear correlation coefficient is not a transitive measurement. That is, if *X* is correlated with *Y*, and *Y* is a variable correlated with *Z*, this does not necessarily imply that between *X* and *Z* there is a statistical relationship of linear type.

2. Factors that impact on the size and interpretation of the linear simple correlation coefficient *r*

The main factors that impact on the size and interpretation of the linear simple correlation coefficient *r* are: distribution shape, size of empirical data sample, outliers, restriction of the empirical data amplitude, non-linearity, measurement errors and the third variable (or several). All statistics textbooks specify that the values of Pearson linear correlation coefficient belong to the interval [-1, 1]. But marginal values are reached only if the distributions of factorial variables *X* and *Y* are symmetric and have roughly the same shape. The absolute value of *r* is lower than 1 if the distributions of variables *X* and *Y* have different shapes.

[3] Kahane, H, 2001, *Regression Basics*. SAGE, London

The absolute value of *r* is lower than 1 if the distributions of variables *X* and *Y* have different shapes. If both distributions of the correlative pair variables do not share the same shape, the increase of the factorial variable *X* shall not be always accompanied by the increase of variable *Y* (for the positive relationship) and decrease of variable *Y* (for the negative relationship). According to Caroll (1961), the less similar the distributions shape from one variable to another, the lower the maximal value of Pearson correlation coefficient. It is also impossible in practice to obtain a -1 correlation.

The size of the empirical data sample may also impact on the accuracy of the correlation intensity estimation, especially when the sample size is small, the standard error for *r* increases in this case. According to Wishart (1931), for instance when the sample size is of 20 statistical units, about 95% of the correlation coefficients have values ranging between [-0.47, +0.47] and we shall be tempted to consider the relationship between the analysed variables as having an average intensity; but if the sample has 102 registered statistical units, 95% of the correlation coefficients shall have values ranging between [-0.2, +0.20], therefore the relationship has a weak intensity[4]. The conclusion to be drawn up is that we must be very careful when interpreting the Pearson correlation coefficient, calculated on the basis of a small sample of empirical data.

The "outliers" are extreme values of the empirical data that may drastically affect the value of the Pearson correlation coefficient, especially if the data sample size is smaller. Outliers may be found in the empirical data that constitute a variable, two or both.

For instance, suppose we have the following example. a sample made of 25 teachers for whom we wish to analyse the possible existence between seniority (factorial variable *X*) and wage (factorial variable *Y*). The linear correlation coefficient, as the relationship between the two variables is linear according to the graphical examination, for the correlative pair *(X, Y)* is of 0.705 (calculation made by means of SPSS 8.0), coefficient that implies the existence of a strong correlation. If we remove only case 7, the linear correlation coefficient is of 0.533, average correlation, and if we remove both abnormal cases, the linear correlation coefficient is of 0.939, a very strong correlation, almost functional.

As one can notice in this example, the outliers existence especially in the case of small samples, may very easily impact on both size and direction of Pearson linear correlation coefficient. In general, big samples offer the possibility to determine the linear simple correlation coefficient more accurately, as it is less affected by the outliers.

Amplitude limitation and non-linearity are also two situations that may affect the quantification of the correlation by means of Pearson coefficient. The limitation of amplitude for the statistical data may occur when the methods for variables measurement are not responsive enough to include all their characteristics. For instance, in case of social surveys, subjects would generally refuse to answer delicate questions about alcohol or drug abuse for example. Therefore, the distribution of such variables is automatically truncated, asymmetric.

The restriction of the empirical data amplitude may also occur when researchers choose relatively homogeneous samples for their surveys. This type of amplitude restriction is called "incidental selection" in the speciality literature (Glass & Hopkins, 1996). Subjects in the homogeneous samples have the same common set of statistical characteristics (for example personality, education, living standard, geographical localization etc.) the correlation coefficient may either increase or decrease, according to the empirical data type and amplitude, generally tending to decrease if the amplitude is restricted (smaller) and the sample is homogeneous.

[4] Wishart, J. (1931), The Mean and Second Moment Coefficient of the Multiple Correlations Coefficient in Samples from a Normal Population. *Biometrika*, 22, pp.353-361.

In case of non-linearity, the use of Pearson correlation coefficient is a wrong decision, its values leading to misinterpretations. The visualization of distribution of pairs (x_i, y_i) is first recommended by means of graphical representation. If we are dealing with a non-linear relationship, the use of the correlation ratio is recommended, or we may change the data or use the polynomial regression (Bobko, 1995; Pedhazur, 1973).

The variables are often correlated statistically, but in reality there is no causal relationship between them. This relationship bears the name of "spurious correlation" in the literature. The occurrence of such a correlation is ascribed to the influence of one or several so-called "the third variable". For instance, the economic literature often mentions the positive relationship between age and satisfaction generated by one person's job. Is it possible that this spurious correlation be generated the relationship between these two variables and a third one, such as the period of time one person stays employed in a firm? Surveys show that people satisfied with their job have a longer stationary duration in an organization.

Therefore the satisfaction generated by the job influences this third variable. Similarly, if we move from Northern Europe towards South, the proportion of Romano-Catholic religion among inhabitants increases. In the same time, a decrease of inhabitants' average height is noted. If we like to detect a correlation between population's height and proportion of Romano-Catholic religion we would find a convincing enough negative correlation. But this is an illusive correlation, as the population's height depends on totally different factors. However the association does not necessarily involve causality as well.

3. Special cases of Pearson linear correlation coefficient

We shall present in this section special indicators used to quantify the relationship between two variables of a correlative pair: Phi coefficient (ϕ), biserial coefficient (r_{bis}), tetrachoric coefficient (r_{tet}) and eta coefficient (η).

Phi coefficient **(ϕ)** is a special case of Pearson linear correlation coefficient, used when both variables of the correlative pair are qualitative (dichotomic). For instance, let's analyse the possible relationship between sex, factorial variable *X*, and acceptance to the Faculty of Economics, resulting variable, *Y*. We shall analyze two samples of empirical data, one of them made of 30 female persons, of which 10 are accepted persons and a sample made of 40 male persons, of which 25 are accepted. The empirical data are shown in table 1, their decoding being made as follows: 2 for female sex, 1 for male sex 1 for acceptance, 0 for rejection.

Table 1: *Acceptance to the Faculty of Economics for male and female candidates.*

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
2	1	2	0	2	0	1	1	1	1	1	1	1	0
2	1	2	0	2	0	1	1	1	1	1	1	1	0
2	1	2	0	2	0	1	1	1	1	1	1	1	0
2	1	2	0	2	0	1	1	1	1	1	1	1	0
2	1	2	0	2	0	1	1	1	1	1	1	1	0
2	1	2	0	2	0	1	1	1	1	1	0	1	0
2	1	2	0	2	0	1	1	1	1	1	0	1	0
2	1	2	0	2	0	1	1	1	1	1	0	1	0
2	1	2	0	2	0	1	1	1	1	1	0	1	0
2	1	2	0	2	0	1	1	1	1	1	0	1	0

Based on Pearson's formula *r*, we shall calculate phi correlation coefficient (ϕ) among sex, as factorial variable, and rejection, as resulting variable. We have:

$$\sum XY = 45,\ \sum X = 100,\ \sum X^2 = 160,\ \sum Y = 35,\ \sum Y^2 = 35.$$

$$\phi = \frac{45 - \frac{160 * 35}{70}}{\sqrt{160 - \frac{1000}{70}}\sqrt{35 - \frac{35^2}{70}}} = -0.69$$

This negative correlation suggests that there are differences between acceptance of male and female persons, the former having more acceptances than the latter. Applying the bivariate statistical test *t* to see if there is statistical significance, we have: t(68)=-2.50 with the probability p<0.05, therefore there is statistical significance for the relationship between sex and acceptance. The same result is obtained using the test χ^2. Because $\chi^2 = n\phi^2$ *cu* $df = 1$, we will obtain the observed value of χ^2 equal to 5.89, which is higher than the critical value of χ^2 of 3.84, at a significance threshold $\alpha = 0.05$.

The biserial coefficient (r_{bis}) is used when the variable *X* has a normal distribution but it is artificially measured as a qualitative variable (for example passed/failed or efficient/inefficient) in its relationship with a continuous resulting variable *Y* (for example capacity test results or labour productivity for a sample of employed persons). In other words, the biserial correlation coefficient is an indicator that estimates which would be the relationship between two variables *X* sand *Y*, if *X* were not artificially converted to a qualitative variable.

Let's consider as a practical example that we want to determine the possible relationship between the operation manner of a device (its productivity, factorial variable *X*) and the operating hours (period of time when one may work with this device, measured in hours/week, resulting variable *Y*). Suppose the sample is made of 20 devices and we shall establish 2 attributes for the operation manner, good, encoded with 2, or weak, encoded with 1, although the variable *X* in fact is still a continuous variable, with a normal distribution. The empirical data corresponding to this example are shown in table 2.

Table 2: *Operation manner (X) and operating hours with this device for one week (Y)*

X	Y	X	Y
1	18	2	10
1	10	2	7
1	16	2	12
1	20	2	16
1	16	2	13
1	16	2	7
1	12	2	7
1	15	2	14
1	12	2	10
1	14	2	12
$\overline{Y}_1 = 14.9,\ s^2{}_1 = 8.989$		$\overline{Y}_2 = 10.8,\ s^2{}_2 = 9.956$	

The biserial correlation coefficient (r_{bis}) is determined using the formula:

$r_{bis} = \frac{\overline{Y}_2 - \overline{Y}_1}{s_Y}(\frac{n_2 n_1}{\lambda(n_1+n_2)^2})$, where $\overline{Y}_2$ *si* $\overline{Y}_1$ are the means of the values registered for variable *Y* for samples 1 and 2, n_1 *si* n_2 represents the samples size and s_Y is the standard deviation for variable Y; λ represents the normal distribution order at $n_2/(n_1+n_2)$ subjects percentage in sample 2 (the values corresponding to λ are listed). Therefore we shall have:

$$r_{bis} = \frac{10.8-14.9}{3.66}(\frac{10*10}{0.33989*(10+10)^2}) = -0.70 .$$

If we quantified the relationship between *X* and *Y* using the biserial-point coefficient previously described we would have: $r_{pb} = -0.57$. The relationship between r_{bis} and r_{pb} can be mathematically presented as:

$r_{bis} = r_{pb}\sqrt{\frac{n_2 n_1(n_1+n_2-1)}{\lambda^2(n_1+n_2)^3}}$. Because $\sqrt{\frac{n_2 n_1(n_1+n_2-1)}{\lambda^2(n_1+n_2)^3}} \geq 1.25$, r_{bis} is always higher than r_{pb} (Glass & Hopkins, 1996). In certain situations, if for example the distribution of a continuous variable is bimodal or platicurtic, r_{bis} it may be even higher than 1 (McNemar, 1969).

The tetrachoric correlation coefficient, marked with r_{tet}, is used to determine the intensity of the relationship between two continuous variables *X* and *Y*, yet both artificially turned into qualitative variables (dichotomic ones). For instance, changing the example previously shown and artificially turning the continuous resulting variable *Y* into a qualitative variable, as short operation time (less than 12.5 hours/week) and long operation time (more than 12.5 hours/week), we have a case where the intensity of the relationship between the two variables is determined using the tetrachoric coefficient r_{tet}. The corresponding data are shown in table 3.

Table 3: *Operation manner (X) and operation time (Y)*

	SHORT (Y=0)	LONG (Y=1)	TOTAL
Good (X=1)	6=a	4=b	10=(a+b)
Weak (X=0)	4=c	6=d	10=(c+d)
Total	10=(a+c)	10=(b+d)	20=a+b+c+d

The tetrachoric correlation coefficient may be calculated using the formula:

$$r_{tet} = \frac{bc-ad}{\lambda_X \lambda_Y n^2}, \quad (4)$$

where *n* represents the sample size, λ_X is the standard distribution order to *(a+b)/(a+b+c+d)* statistical units proportion for which *X=1*, λ_Y is normally standardized distribution order to *(b+d)/(a+b+c+d)* statistical units proportion for which *Y=1*. Therefore we shall have:

$$r_{tet} = \frac{16-36}{0.3989*0.3989*20^2} = -0.32.$$

The tetrachoric correlation coefficient is not a faithful measurement of the intensity of the relationship between two variables except for large samples of about 400 or more records. (Glass & Hopkins, 1996). The continuous variables should not be artificially turned into qualitative variables unless there are good reasons to do so. Both the biserial and the tetrachoric correlation coefficients are rarely used in practice, as their use requires particular attention, because they estimate the intensity of a hypothetical correlation.

Eta correlation coefficient η is an indicator of intensity of association between a multichotomic variable *X* with *n* distinct categories and a variable *Y* measured on variation intervals. Unlike the correlation coefficients previously presented, η describes only the intensity of the relationship between variables, its direction lacking importance, since the categories of the multichotomic variable *X* do not reflect the existence of any sequential order.

The eta correlation coefficient may be also used to describe a curvilinear relationship between a qualitative variable and a variable measured on variation intervals. The eta correlation coefficient is usually used within the analysis of variance ANOVA. For instance, if we'd like to know if there is any difference between the growth rates averages of three types of macroeconomic indicators, net domestic product, net national product and net income. We shall monitor the growth rates for 10 cases of each type of macroeconomic indicator, the results can be found in table 4.

Table 4: *The growth rates of the three macroeconomic indicators*

NET DOMESTIC PRODUCT	NET NATIONAL PRODUCT	NET INCOME
2.5	3.2	3.5
2.6	3.4	3.2
2.7	2.6	3.0
3.2	3.2	3.0
2.8	3.9	3.6
2.4	2.7	2.9
2.1	3.1	3.3
2.0	2.9	4.1
2.5	3.4	3.2
2.2	2.9	3.5
$n_{PIN} = 10$ $\overline{Y}_{PIN} = 2.50$ $s_{PIN} = 0.36$	$n_{PNN} = 10$ $\overline{Y}_{PNN} = 3.13$ $s_{PNN} = 0.38$	$n_{VN} = 10$ $\overline{Y}_{VN} = 3.33$ $s_{VN} = 0.36$

The results of analysis of variance show there are significant differences between the growth rates averages for the three macroeconomic indicators, and a significant association between the type of macroeconomic indicators and the growth rate. This association may be

measured using eta correlation coefficient: $\eta = \sqrt{\frac{SS_{\text{int}\, regrupe}}{SS_{total}}}$. Taking into account the results obtained using the analysis of variance, we shall have: $SS_{\text{int}\, regrupe}$=3.753, $SS_{total} = 7.375$ and $\eta = 0.71.$, therefore different from zero ($F = 13.99, p < 0.05$).

References:

Bobko P., (1995), *Correlation and Regression: Principles and Applications for Industrial Organizational Psychology and Management.* New York: Mc.Graw-Hill.

Cahan S., (1987), *On the Interpretation of the Product Moment Correlation Coefficient as a Measure,*" unpublished manuscript, The Hebrew University, School of Education, Jerusalem, Israel.

Carroll J. B., (1961), *The nature of the data, or how to choose a correlation coefficient. Psychometrik*a, 26, 347-372.

Fisher R., (1970), Statistical *Methods for Research Workers. 14th edition*. Oliver and Boyd, Edinburgh.

Glass G.V., & Hopkins K.D., (1996), *Statistical methods in education and psychology*, Boston: Allyn and Bacon.

Kahane H., (2001), *Regression Basics*. SAGE, London

Ozer D. J., (1985), *Correlation and the Coefficient of Determination*, Psychological Bulletin*,* 97, 307-315.

Rodgers J. L., and Nicewander W. A., (1988), *Thirteen Ways to Look at the Correlation Coefficient*, The American Statistician, 42, 59-66.

Wishart J., (1931), *The Mean and Second Moment Coefficient of the Multiple Correlations Coefficient in Samples from a Normal Population*. Biometrika, 22, pp.353-361.